Descriptive Statistics (Quantitative)

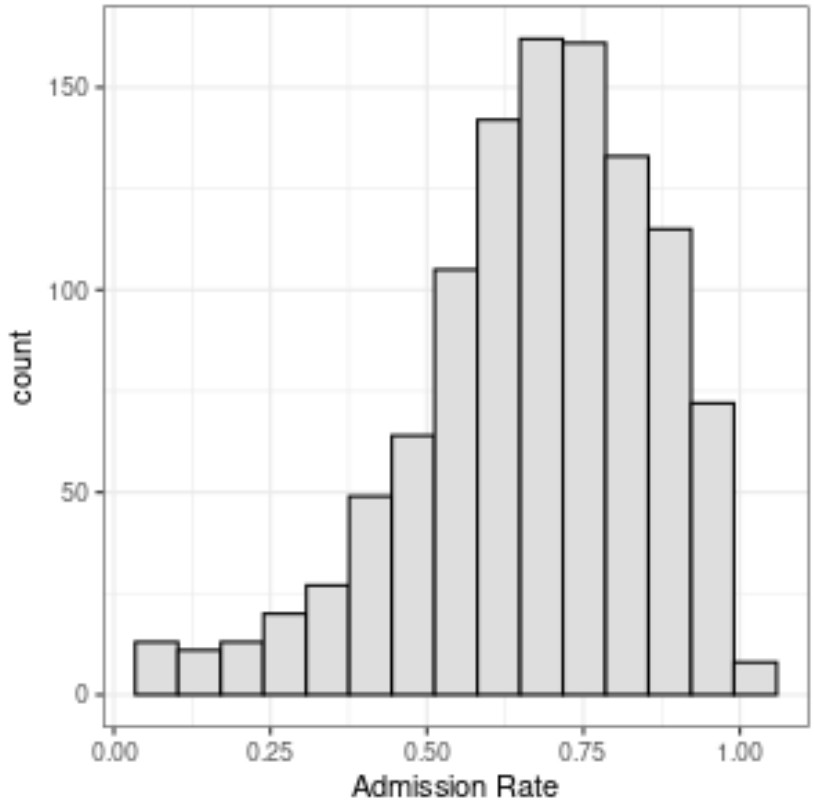
Grinnell College

February 9, 2024

Review

Qualitative Distributions

- Shape
- Center
- Spread



Measures of Centrality

There are two ways to measure centrality in quantitative variable:

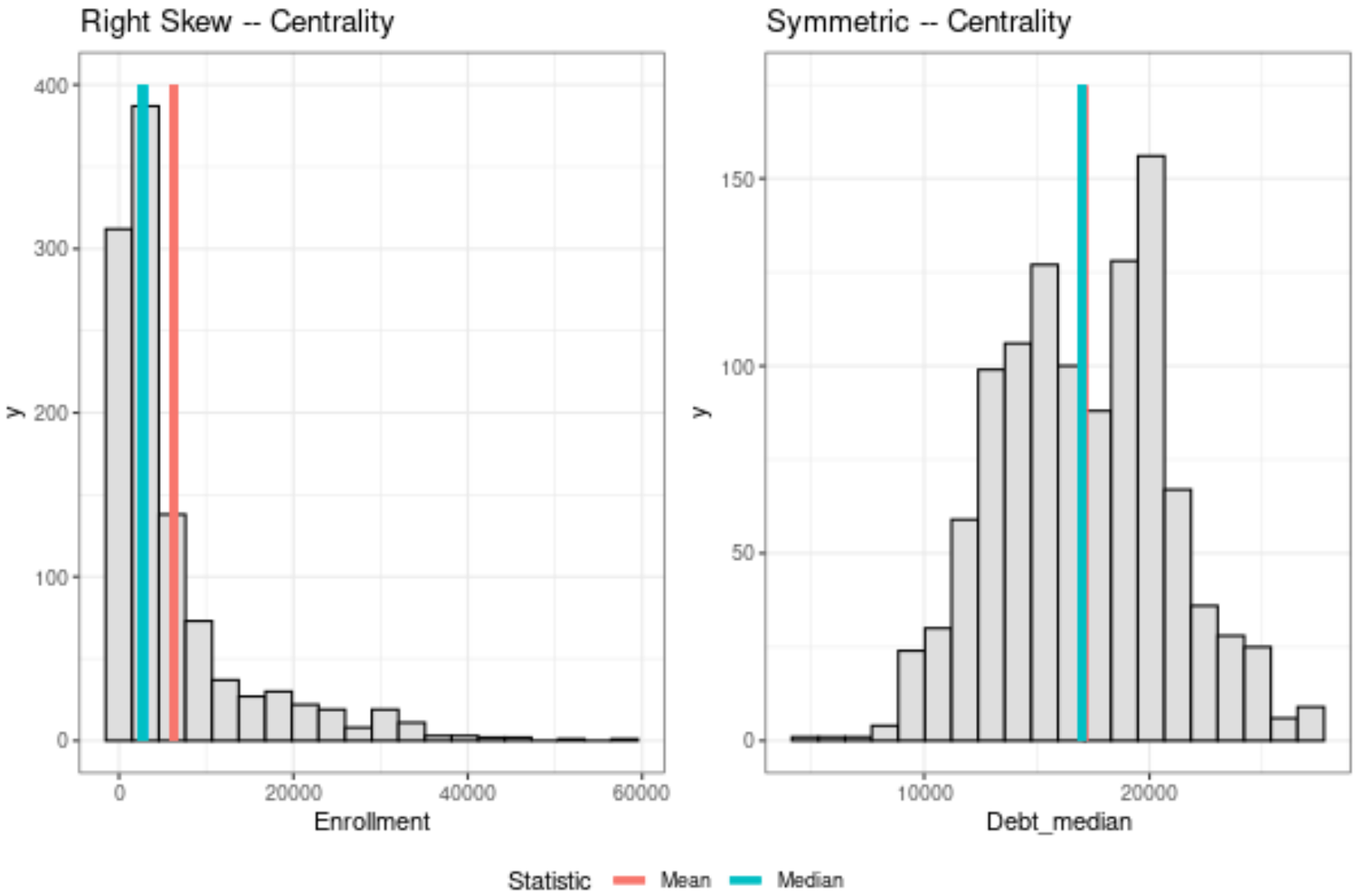
- **Mean** – the arithmetic average of a variable

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Median** – the middle value of the data if arranged from smallest to largest

The median is called a **robust** statistic as it tends to not be influenced by outliers; this is *not* true for the mean

Measures of Centrality



Measures of Spread

Important ways to summarize spread:

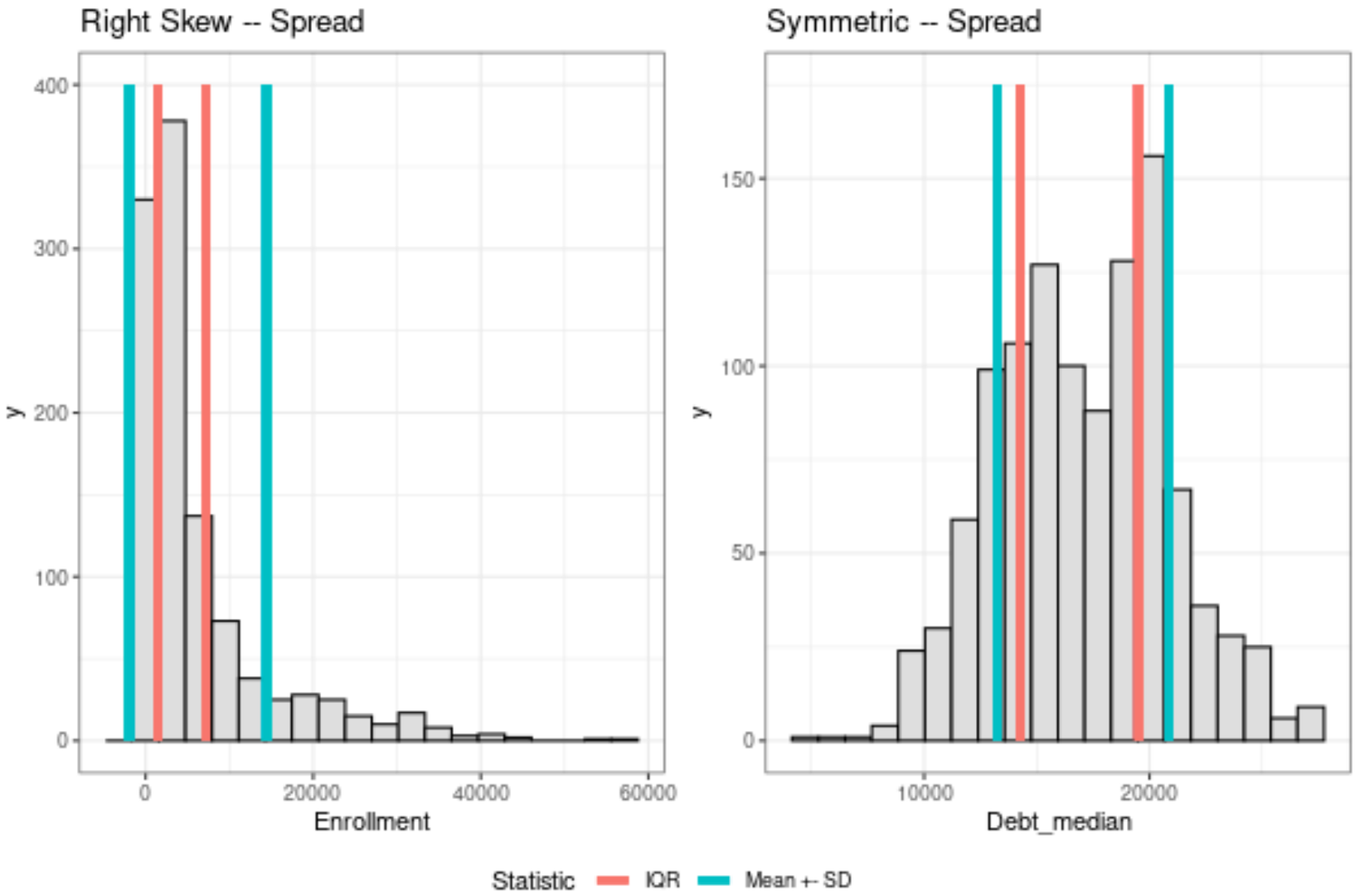
- **Standard deviation** – the average deviation (distance) of individual observations from the mean value

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

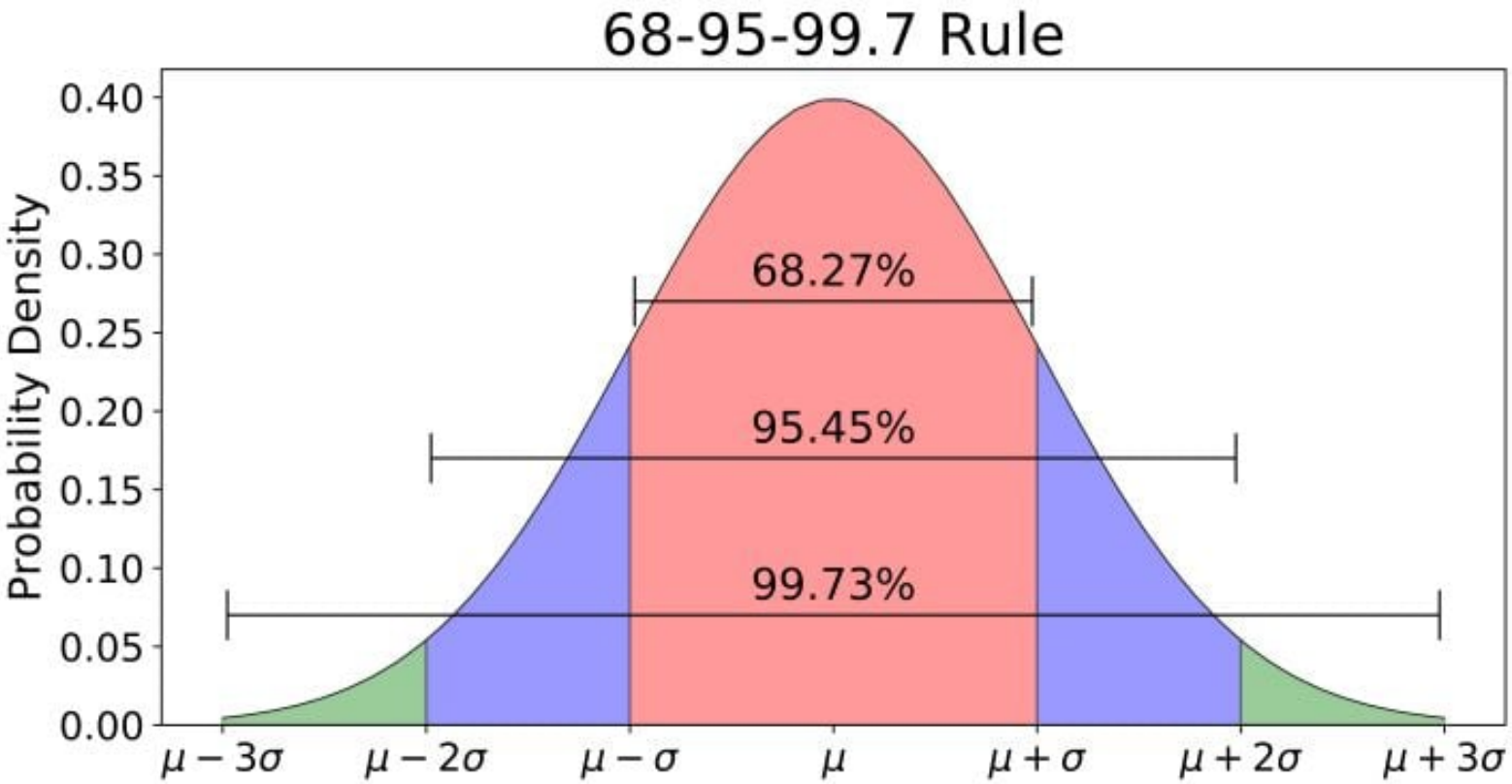
- **Range** – the difference between smallest and largest values
- **Interquartile Range (IQR)** – the difference between the 75^{th} quartile and the 25^{th} quartile

The standard deviation and range are *greatly* influenced by outliers, whereas the IQR is considered *robust*

Measures of Spread



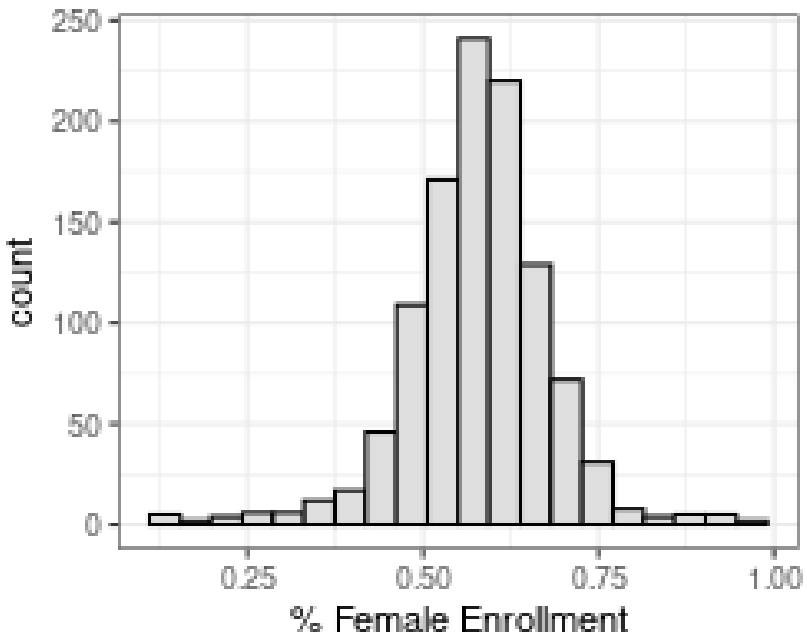
68-95-99 Rule

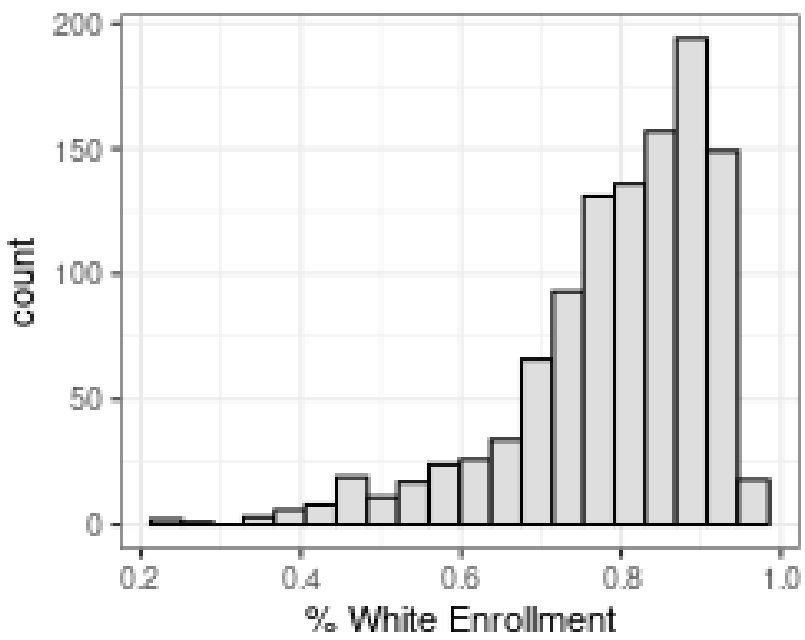

68-95-99.7 Rule
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
68.27%
95.45%
99.73%
$\mu - 3\sigma$
$\mu - 2\sigma$
$\mu - \sigma$
μ
$\mu + \sigma$
$\mu + 2\sigma$
$\mu + 3\sigma$

Practice

For each of the following variables visualized below:

1. Determine approximate mean and median and which should be larger. How do you know?
2. Decide whether standard deviation or IQR is more appropriate for describing variability





Conditional Statistics

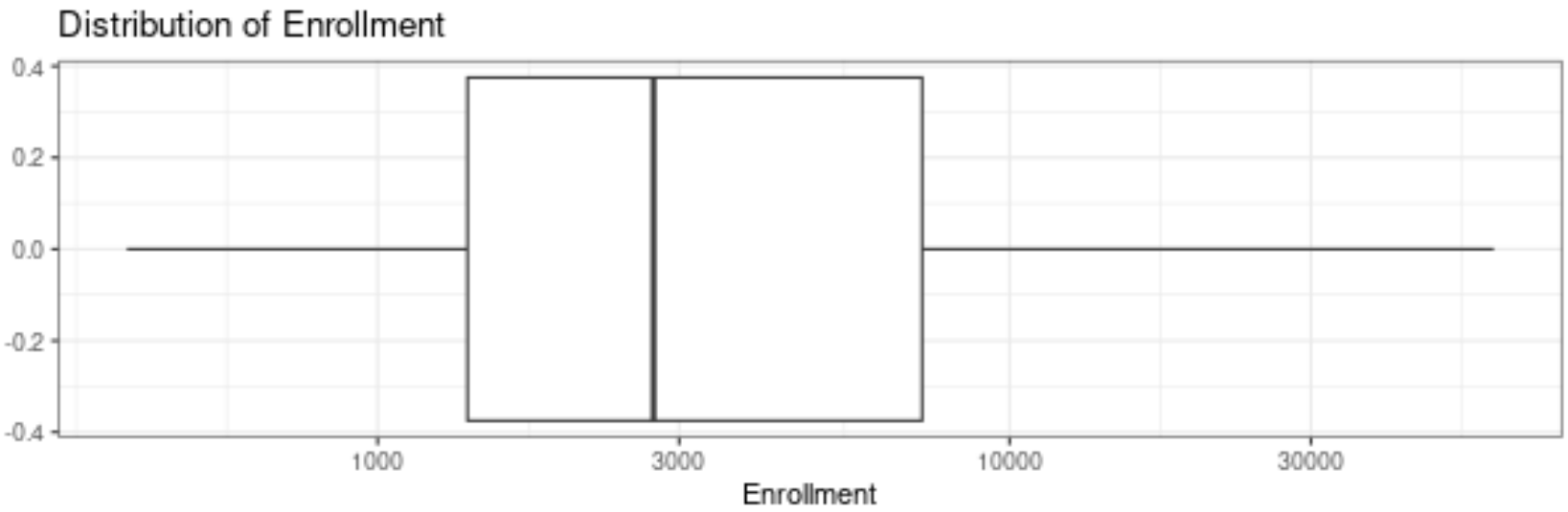
Consider the five-figure summary for our Enrollment variable

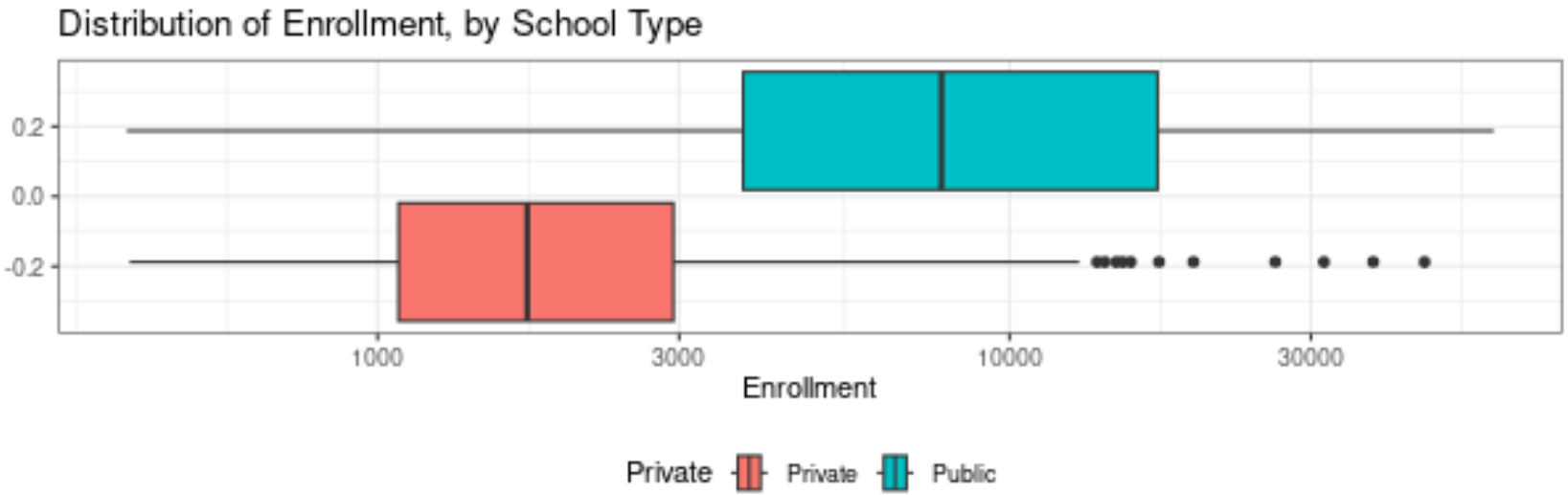
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
401	1388	2733	6241	7272	58392

Note that we typically refer to measures of centrality when discussing association

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Private	405	1079	1725	2720	2938	45370
Public	401	3788	7803	11325	17152	58392

Conditional Statistics





Digression – z-scores

A common technique when using quantitative variables involves *standardizing* our values to create **z-scores**:

$$z_i = \frac{x_i - \overline{x}}{s_x}$$

The transformed variable z will have the same observations as the original variable x, but now with a mean of 0 and a standard deviation of 1

z-scores

For example, in our college dataset, the average ACT Median is $\overline{x} = 23.58$, with a standard deviation of $s_x = 3.55$

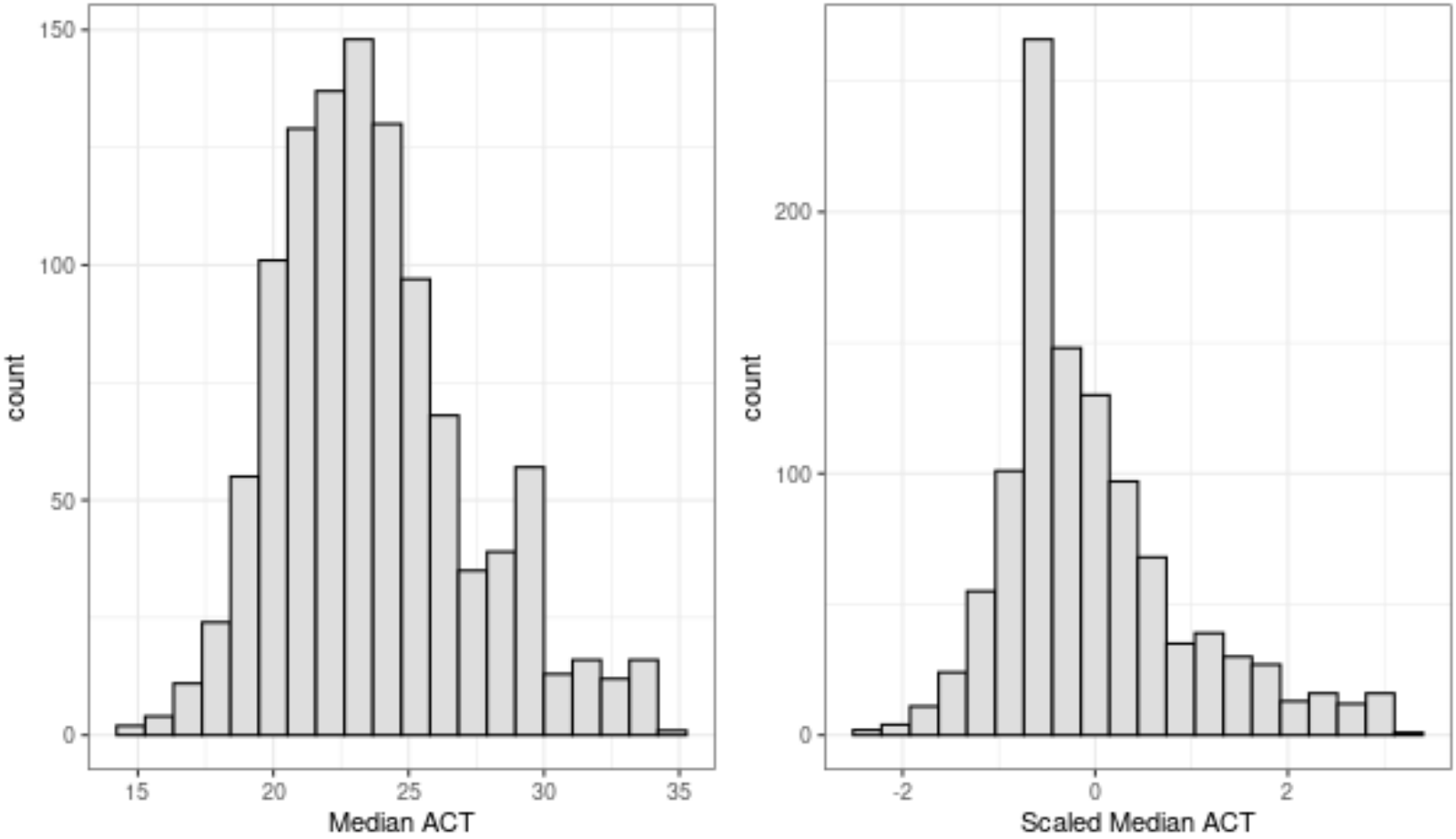
- Grinnell College has a median ACT of 32
- We can calculate its standardized value as:

$$z_{Grinnell} = \frac{32 - 23.58}{3.55} = 2.37$$

- This indicates that the median ACT at Grinnell College is 2.37 standard deviations above the average

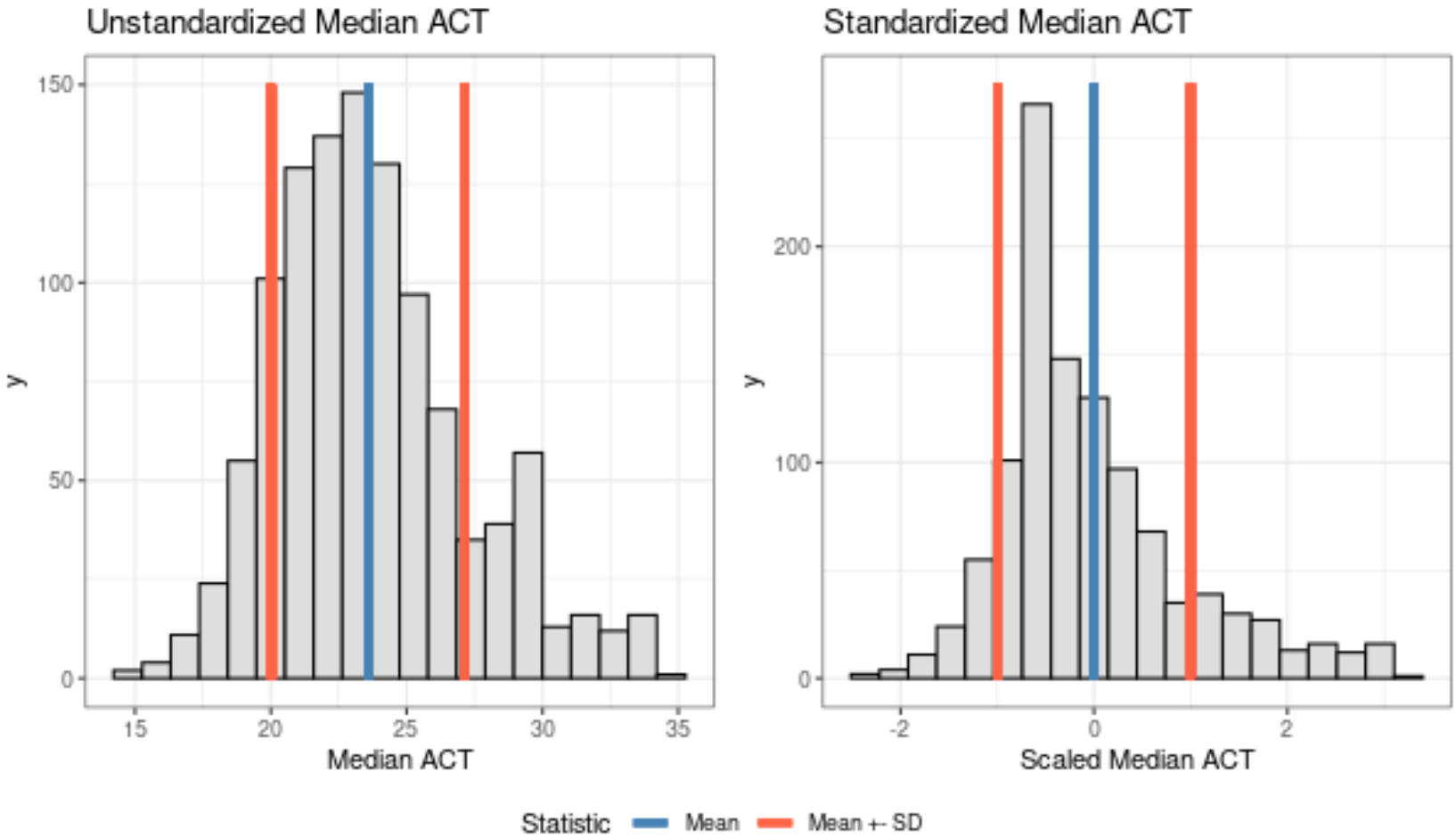
Transformation

Values are either squeezed or stretched, but their relative positions stay the same



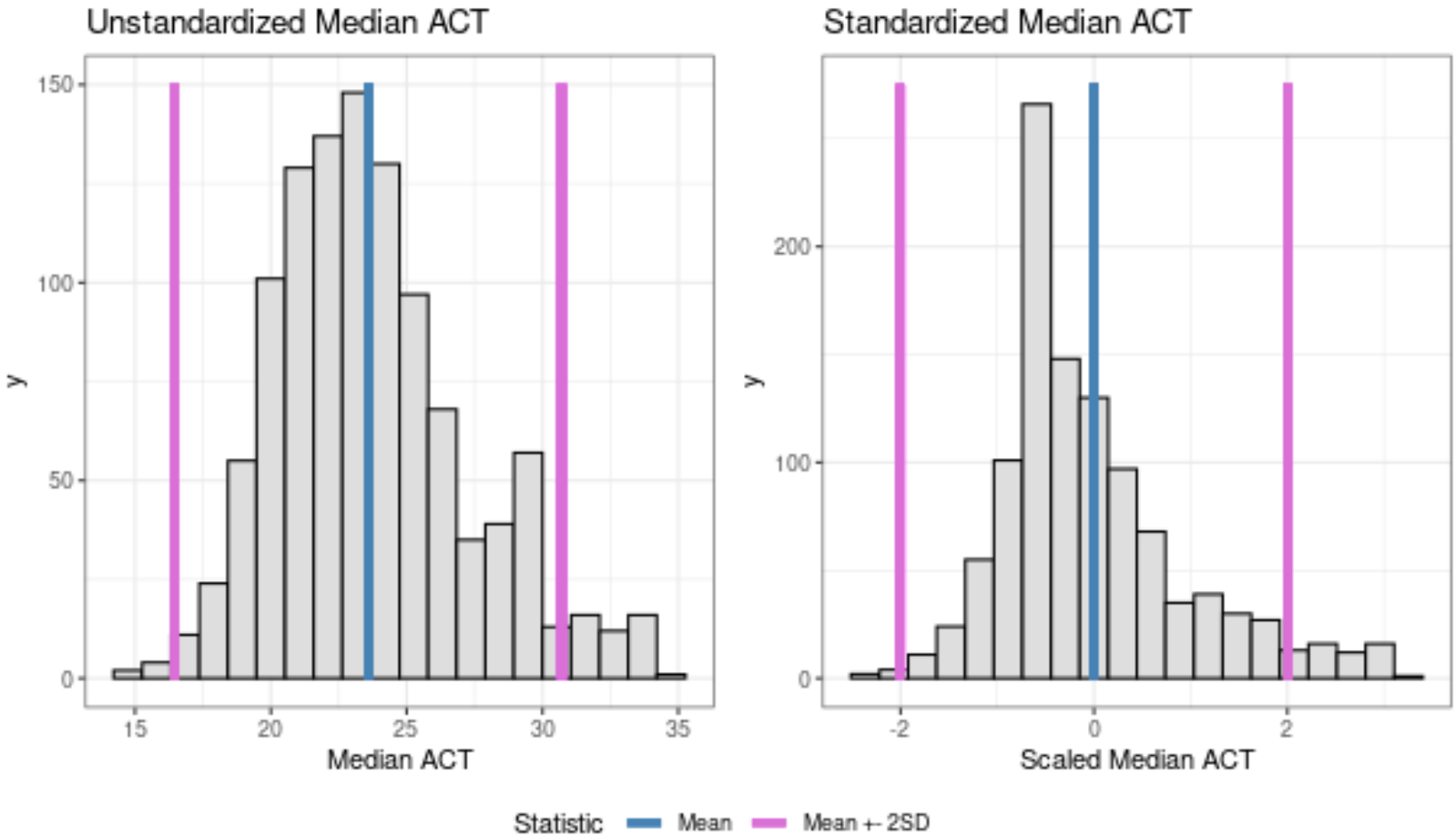
Transformation

Mean and one standard deviation

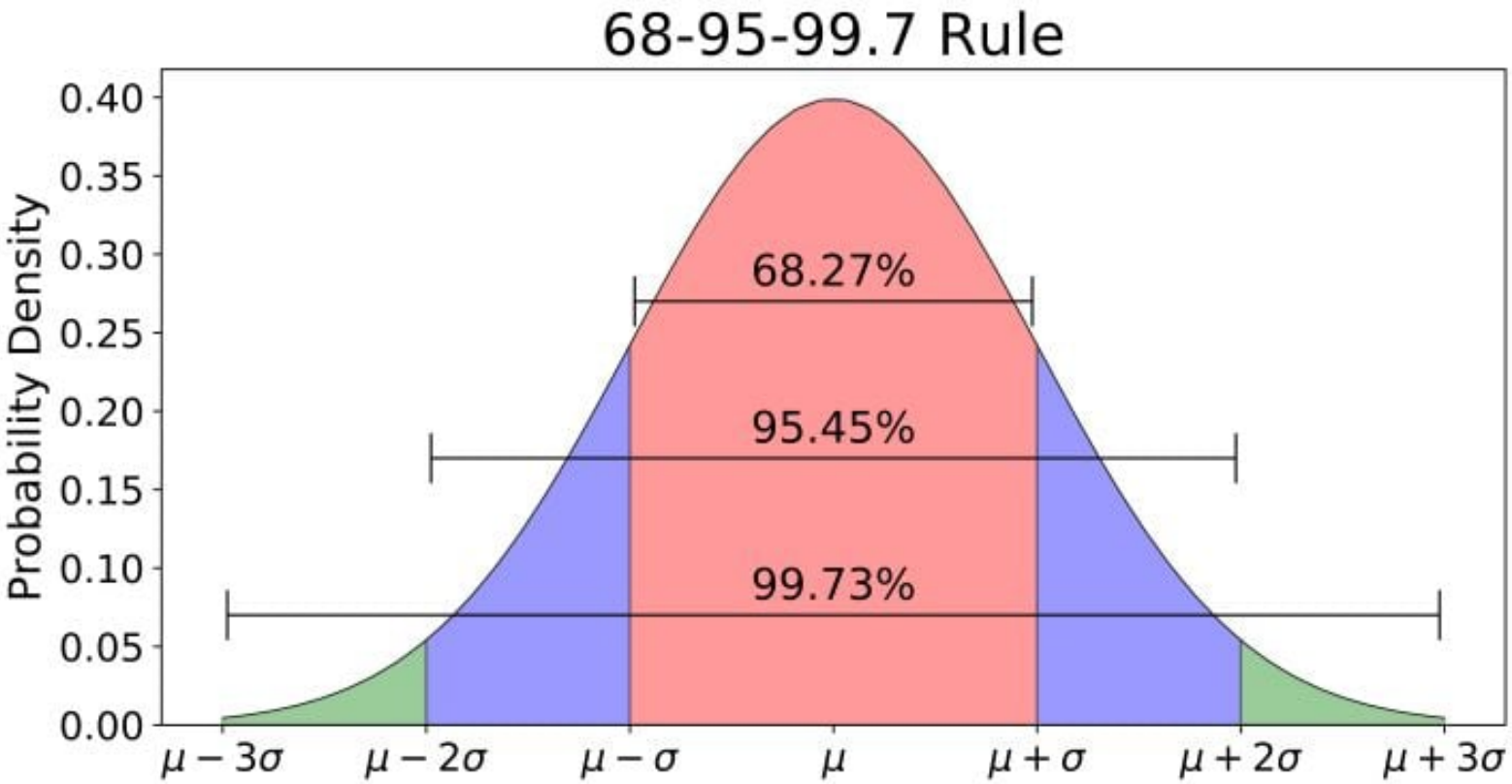


Transformation

Mean and two standard deviations



68-95-99 Rule Again



Review

- Measures of centrality
- Measures of spread
- Robust statistics
- Conditional Tables
- Standardization